Universal Adversarial Purification with DDIM Metric Loss for Stable Diffusion

Li Zheng[1], **Liangbin Xie**[1 2], **Jiantao Zhou**[1*], **He YiMin**[1]

[1]University of Macau
[2]Shenzhen Institute of Advanced Technology
yc27908@um.edu.mo, lb.xie@siat.ac.cn, jtzhou@um.edu.mo, mc45058@um.edu.mo

Abstract

Stable Diffusion (SD) often produces degraded outputs when the training dataset contains adversarial noise. Adversarial purification offers a promising solution by removing adversarial noise from contaminated data. However, existing purification methods are primarily designed for classification tasks and fail to address SD-specific adversarial strategies, such as attacks targeting the VAE encoder, UNet denoiser, or both. To address the gap in SD security, we propose Universal Diffusion Adversarial Purification (UDAP), a novel framework tailored for defending adversarial attacks targeting SD models. UDAP leverages the distinct reconstruction behaviors of clean and adversarial images during Denoising Diffusion Implicit Models (DDIM) inversion to optimize the purification process. By minimizing the DDIM metric loss, UDAP can effectively remove adversarial noise. Additionally, we introduce a dynamic epoch adjustment strategy that adapts optimization iterations based on reconstruction errors, significantly improving efficiency without sacrificing purification quality. Experiments demonstrate UDAP's robustness against diverse adversarial methods, including PID (VAE-targeted), Anti-DreamBooth (UNet-targeted), MIST (hybrid), and robustness-enhanced variants like Anti-Diffusion (Anti-DF) and MetaCloak. UDAP also generalizes well across SD versions and text prompts, showcasing its practical applicability in real-world scenarios.

Code — https://github.com/whulizheng/UDAP

Introduction

SD has emerged as a groundbreaking framework in the field of image and video generation (Li et al. 2023; Ramesh et al. 2021; Gafni et al. 2022; Ding et al. 2021), renowned for its ability to synthesize high-quality images and videos. Recent advancements, such as DreamBooth (Ruiz et al. 2023) and LoRA (Hu et al. 2021), have further enhanced the capabilities of SD, positioning it at the forefront of personalized image generation (Chen et al. 2023, 2024; Croitoru et al. 2023). However, all of these methods are vulnerable to adversarial attacks (Van Le et al. 2023; Li et al. 2024; Liang et al. 2023): adding imperceptible noise to the training images can mislead the SD models to generate inaccurate or degraded outputs (shown in the middle panel of Fig. 1).

*Corresponding author

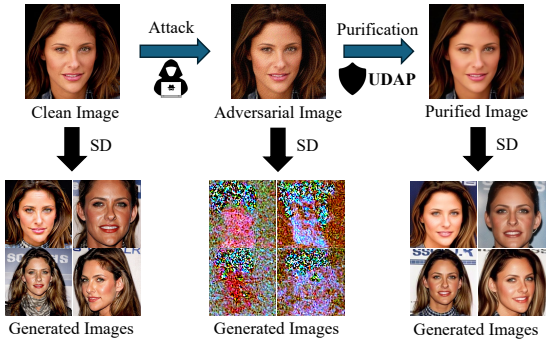

Figure 1: Illustration of the impact of adversarial attacks on SD models and the effectiveness of our proposed UDAP.

In such cases, the models are compromised, resulting in a significant waste of computational resources. Therefore, removing the imperceptible noise in the adversarial images before training the SD models is crucial to defend SD models against such adversarial attacks.

The majority of work on defending neural networks against adversarial attacks has focused on image classification tasks (Baniecki and Biecek 2024; Li, Xin, and Liu 2022; Chakraborty et al. 2021) and can be categorized into adversarial training (Bai et al. 2021; Shafahi et al. 2020, 2019) and adversarial purification (Costa et al. 2024; Liao et al. 2018). In contrast to adversarial training, which is defined to defend against specific attacks that it has been trained on, adversarial purification methods can better defend against previously unseen threats in a plug-and-play manner (Nie et al. 2022). As a result, significant progress has been made in the area of adversarial purification, evolving from the use of autoencoders in MagNet (Meng and Chen 2017), to the application of GANs in Defense-GAN (Samangouei 2018), and more recently, the leveraging of diffusion models (Nie et al. 2022; Zollicoffer et al. 2025; Wang et al. 2024). While adversarial purification techniques for classification tasks have advanced continuously, to the best of our knowledge, there is no research specifically targeting adversarial purification for SD models. Given the evolving nature of adversarial attack techniques targeting SD, there is an urgent need for adversarial purification methods specifically tailored for SD.

Considering that current adversarial attack methods (e.g.,

Anti-DreamBooth (Anti-DB) (Van Le et al. 2023), PID (Li et al. 2024), and MIST (Liang et al. 2023)) targeting SD generate adversarial noise are specifically designed to attack different parts of SD (Truong, Dang, and Le 2024), there is a clear need for a robust method capable of handling various adversarial attacks. Furthermore, in practical scenarios, training sets often contain a mixture of adversarial and clean images. Applying the same adversarial purification to all images would undoubtedly be time-consuming and inefficient, making it impractical for real-world use. Therefore, it is essential to design a dynamic optimization mechanism that can adaptively adjust the strength of purification based on the underlying adversarial noise level in the input images.

These considerations motivate our proposal of a robust adversarial purification method, specifically tailored to defend against various SD adversarial attacks. Although different SD adversarial attack methods target different parts of the SD model, they share a common characteristic: they are optimized in the latent space of the model. This characteristic implies that the latent corresponding to an adversarial image can alter the results of both the forward and inversion processes in SD. DDIM inversion (Song, Meng, and Ermon 2020) is a technique that repeatedly performs inversion on the latent before executing the forward process, which therefore can amplify the distance between the reconstructed image and its corresponding adversarial image. As shown in Fig. 2, we observe that through DDIM inversion, the L_2 loss between a clean image and its corresponding reconstruction is small, whereas the L_2 loss between an adversarial image and its reconstruction is significantly larger. We propose utilizing the advantageous properties of DDIM inversion as a metric loss. Using the latent corresponding to the input image as an initial latent, we perform DDIM inversion on the latent and iteratively optimize it by calculating the L_2 loss between the reconstructed image and the input image. We find that the design of DDIM reconstruction metric loss is simple, yet highly effective, performing well across a variety of adversarial attack methods, SD versions, and diverse prompts. Moreover, to enhance practical efficiency, we further introduce a dynamic optimization strategy by setting a reconstruction loss threshold as a tradeoff of purification strength and computational cost. This dynamic optimization strategy enables adjustment of purification epochs based on the underlying strength level of adversarial perturbations in the input images. As illustrated in Fig. 1, the proposed UDAP effectively purifies images, enabling the SD model trained on these images to produce highly realistic outputs.

Our key contributions are as follows: **1)** As far as we know, UDAP is the first universal adversarial purification method specifically designed for SD models. **2)** Through theoretical analysis, we demonstrate that adversarial samples targeting SD (regardless of their types) will result in significant DDIM reconstruction errors. This is the theoretical foundation of our universal adversarial purification method. **3)** We propose a novel DDIM metric loss to measure the distance between reconstructed and input images. Minimizing this loss optimizes the initial latent into a clean representation, effectively eliminating adversarial noise. Additionally, we introduce a dynamic optimization strategy to adaptively adjust purification epochs, improving efficiency in real-world scenarios. **4)** Through quantitative and qualitative evaluations, our proposed UDAP demonstrates superior purification performance across various adversarial techniques. For instance, when defending against the PID attack on the CelebA-HQ dataset, our proposed UDAP reduces the Face Detection Failure Rate (FDFR) to (0.14), outperforming baselines like GridPure (0.21) by over (33%). Moreover, the dynamic optimization strategy makes UDAP approximately twice faster while maintaining the high purification effectiveness.

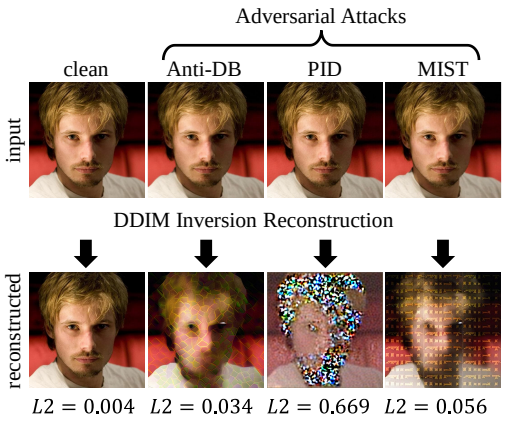

Figure 2: The first row shows the input images and the second row shows their reconstructed images through DDIM inversion reconstruction. Value L_2 means the average L_2 distance between input and reconstructed images.

Related Works

Adversarial Attacks on SD

Although SD is highly capable of personalized image generation, it is vulnerable to adversarial attacks. These attacks insert malicious examples into the training data to corrupt the model's learning process, resulting in distorted or low-quality image outputs. MIST creates adversarial noise targeting both the SD model's encoder and its denoising process for a more thorough disruption. It also watermarks adversarial images, which interferes with image generation and makes the attack difficult to mitigate. Anti-DB simulates the DreamBooth fine-tuning process to dynamically attack the model, making standard data purification defenses less effective. In contrast, PID generates prompt-agnostic adversarial noise to corrupt the VAE encoder, disrupting the model's internal image representations for any given text prompt. Building on these approaches, Anti-DF (Zheng et al. 2025) enhances attack performance by combining adversarial noise with prompt tuning and injecting semantic disruptions tailored for SD models. Meanwhile, MetaCloak (Liu et al. 2024) introduces a meta-learning framework to generate transferable adversarial perturbations through bi-level optimization, marking the advent of a new era of adaptive adversarial attack techniques.

Adversarial Purification

Adversarial purification has emerged as a key defense against adversarial attacks, complementing adversarial training by removing perturbations from inputs before model processing. Early methods used GANs and energy-based models, but recent advances favor diffusion models for their high-quality reconstructions and robustness against unseen attacks (Yoon, Hwang, and Lee 2021). These diffusion models, particularly SD, leverage iterative denoising to purify adversarial inputs, enhancing classifier robustness (Nie et al. 2022; Wang et al. 2022; Lin et al. 2024; Zollicoffer et al. 2025; Zhao et al. 2024). Despite these advancements of adversarial purification, existing methods primarily focus on classification tasks and do not adequately address the unique challenges posed by adversarial attacks on SD models. These defense paradigms are fundamentally misaligned with the demands of generative models. Their primary objective is to preserve output invariance (i.e., a class label) against input perturbations, whereas the goal for a model like SD is to maintain the intricate perceptual quality and semantic coherence of the entire generated output. This core difference in objectives makes the direct application of classification-centric defenses to generative models both ineffective and inappropriate. Consequently, there is a pressing need for a universal adversarial purification framework specifically designed to counteract diverse adversarial techniques targeting SD.

Method

In this section, we first analyze the distinct behaviors of clean and adversarial images under DDIM inversion, which forms the theoretical basis for our method. We then present our proposed UDAP framework, followed by a detailed explanation of the inversion optimization using the DDIM metric loss and the dynamic optimization epochs.

Analyzing DDIM Inversion Reconstruction

Let $\boldsymbol{x}^{\text{adv}}$ be an adversarial example targeting diffusion models, and let $\hat{\boldsymbol{x}}^{\text{adv}}$ denote its reconstruction via DDIM inversion. We introduce a positive constant Q to serve as a lower-bound threshold for the distance, quantifying the minimum impact of a successful adversarial attack.

Proposition 1. *$\|\boldsymbol{x}^{adv} - \hat{\boldsymbol{x}}^{adv}\| \geq Q$ when the timestamp of DDIM inversion process approaches the total time steps T.*

Proof. According to the definition of adversarial examples, for a clean sample $\boldsymbol{x}$, suppose that it exists an adversarial example $\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \boldsymbol{\delta}$ with a small perturbation $\|\boldsymbol{\delta}\|_p \leq \xi$. The noise predictions $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{c})$ and $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{adv}}, t, \boldsymbol{c})$ should satisfy such condition:

$$\|\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{c}) - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{adv}}, t, \boldsymbol{c})\| \geq Q, \tag{1}$$

where ϵ is a well trained diffusion model with parameters $\boldsymbol{\theta}$, $\boldsymbol{c}$ is the text embedding of input prompt and Q denotes a clearly perceptible distance.

According to (Song, Meng, and Ermon 2020), the DDIM inversion process $\boldsymbol{x}_t = q_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{c})$ can be defined as $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_+\sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t$ represents the predefined parameters of DDIM. We can have:

$$\begin{aligned}\|\boldsymbol{x}_t - \boldsymbol{x}_t^{\text{adv}}\| = \|\sqrt{\bar{\alpha}_t}(\boldsymbol{x} - \boldsymbol{x}^{\text{adv}}) + \\ \sqrt{1-\bar{\alpha}_t}\left(\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{c}) - \epsilon_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{\text{adv}}, t, \boldsymbol{c}\right)\right)\|\end{aligned} \tag{2}$$

By the definition of adversarial sample $\boldsymbol{x}_t^{\text{adv}}$ and substituting (1) into (2), we have $\|\boldsymbol{x}_t - \boldsymbol{x}_t^{\text{adv}}\| \geq \|\sqrt{\bar{\alpha}_t}\delta + \sqrt{1-\bar{\alpha}_t}Q\|$. When $t \to T$ (total time steps of DDIM), by the definition of DDIM, $\bar{\alpha}_t \to 0$, thus, we can have:

$$\|\boldsymbol{x}_t - \boldsymbol{x}_t^{\text{adv}}\| \geq \|\sqrt{\bar{\alpha}_t}\delta + \sqrt{1-\bar{\alpha}_t}Q\| \approx Q. \tag{3}$$

Moreover, by the process of DDIM inversion, both $\boldsymbol{x}_t$ and $\boldsymbol{x}_t^{\text{adv}}$ should eventually follow Gaussian distributions through iterative noise injection.

Similarly, we define the DDIM denoise process as $\hat{\boldsymbol{x}}_0 = p_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c})$. Then, by the reversibility of DDIM (Song, Meng, and Ermon 2020) and given that $\boldsymbol{x}_t^{\text{adv}}$ follows a Gaussian distribution, there should exist a normal sample $\boldsymbol{x'}$ such that $\boldsymbol{x}_t^{\text{adv}} = q_{\boldsymbol{\theta}}(\boldsymbol{x'}_0, t, \boldsymbol{c})$. By definition, we have $\hat{\boldsymbol{x}}_0^{\text{adv}} = p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{\text{adv}}, t, \boldsymbol{c}) = p_{\boldsymbol{\theta}}(q_{\boldsymbol{\theta}}(\boldsymbol{x'}_0, t, \boldsymbol{c}), t, \boldsymbol{c}) = \hat{\boldsymbol{x}}_0^{'}$, where $q_{\boldsymbol{\theta}}(\boldsymbol{x}_0, t, \boldsymbol{c})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c})$ are inverse functions of each other and $\hat{\boldsymbol{x}}_0^{'}$ is the DDIM inversion reconstructed $\boldsymbol{x}_0^{'}$. Thus, by the reversibility of DDIM, we have:

$$\boldsymbol{x'}_0 \approx \hat{\boldsymbol{x}}_0^{'} \approx \hat{\boldsymbol{x}}_0^{\text{adv}}. \tag{4}$$

Under the assumption that $\epsilon_{\boldsymbol{\theta}}$ is properly trained, it should obey a Lipschitz continuity condition with L_t (where L_t denotes the Lipschitz constant at timestamp t), so we have:

$$\frac{\|q_{\boldsymbol{\theta}}(\boldsymbol{x}_0, t, \boldsymbol{c}) - q_{\boldsymbol{\theta}}(\boldsymbol{x'}_0, t, \boldsymbol{c})\|}{\|\boldsymbol{x}_0 - \boldsymbol{x'}_0\|} = \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_t^{\text{adv}}\|}{\|\boldsymbol{x}_0 - \boldsymbol{x'}_0\|} \leq L_t. \tag{5}$$

Noting that $\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \delta \approx \boldsymbol{x}$ (δ is a very small perturbation) and substituting (3) and (4) into (5), we get:

$$\|\boldsymbol{x}^{\text{adv}} - \hat{\boldsymbol{x}}^{\text{adv}}\| \geq \frac{Q}{L_t}. \tag{6}$$

As $t \to T$, according to the Lipschitz continuity of diffusion models, we have $L_t \ll 1$ (Yang et al. 2024). So (6) can be rewritten as:

$$\|\boldsymbol{x}^{\text{adv}} - \hat{\boldsymbol{x}}^{\text{adv}}\| \geq \frac{Q}{L_t} \gg Q. \tag{7}$$

Therefore, the distance between $\boldsymbol{x}^{\text{adv}}$ and $\hat{\boldsymbol{x}}^{\text{adv}}$ is much larger than Q, where Q denotes a clearly perceptible distance by the definition of adversarial examples. □

Whether an attack targets the UNet or the VAE, it ultimately introduces errors in the latent space that propagate to the final image reconstruction during DDIM inversion. As can be seen from Fig. 2, the experimental results further verify this conclusion. As demonstrated in this figure, while clean images maintain accurate reconstruction quality (with low L_2 distance 0.004), all adversarial images—regardless of their attack types (Anti-DB for UNet-targeted attacks, PID for VAE-targeted attacks, or MIST for mixed attacks)—show significant differences (with high L_2 distances 0.034, 0.669 and 0.056) between the reconstructed images and the original input. These observations highlight that adversarial images can be distinguished from clean images by DDIM inversion reconstruction. As a result, we integrate DDIM inversion reconstruction into our optimization process which forms the foundation of the proposed UDAP.

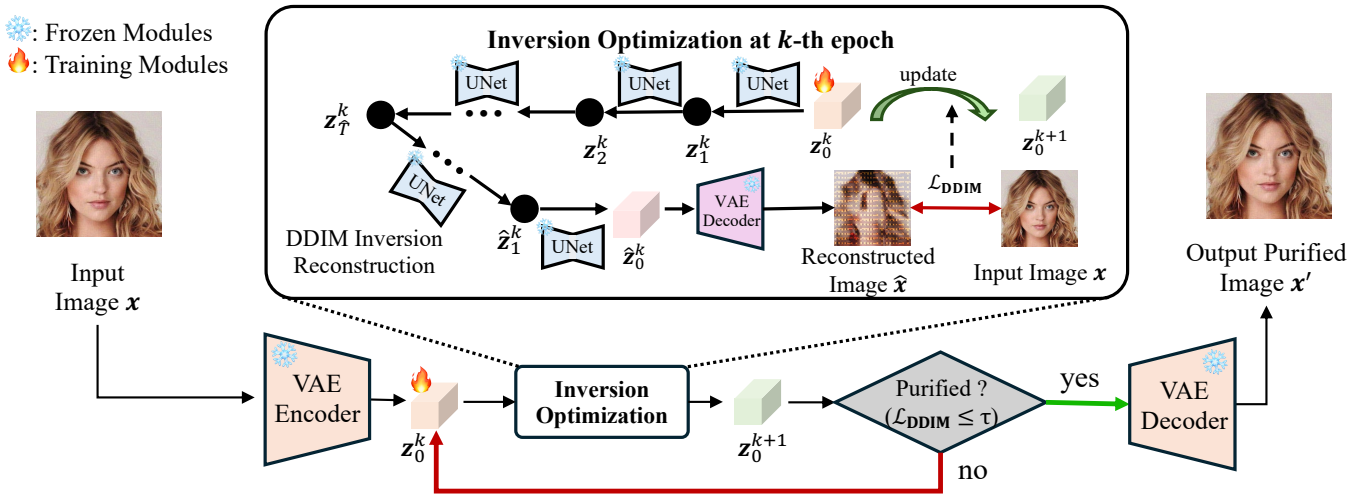

Figure 3: The overall framework of UDAP. The input image is first encoded into a latent space representation, where it undergoes iterative purification through inversion optimization. This optimization is guided by our proposed DDIM metric loss, $\mathcal{L}_{\text{DDIM}}$. Once the latent representation has DDIM metric loss that is less than τ (deemed as sufficiently purified), it is decoded back into the image space using a VAE decoder, resulting in the final purified image.

Overview Framework

The overall framework of UDAP is illustrated in Fig. 3. The process begins with an input image $\boldsymbol{x}$, which is first encoded by the VAE encoder of SD to obtain the initial latent representation $\boldsymbol{z}_0^0 = \mathcal{E}(\boldsymbol{x})$. At the k-th epoch, the latent $\boldsymbol{z}_0^k$ undergoes inversion optimization to produce $\boldsymbol{z}_0^{k+1}$. The value of $\mathcal{L}_{\text{DDIM}}$ at each epoch is also used to determine whether $\boldsymbol{z}_0^{k+1}$ has been sufficiently purified. This is achieved by comparing $\mathcal{L}_{\text{DDIM}}$ with a predefined threshold τ, which is a tradeoff of purification strength and computational cost. If $\mathcal{L}_{\text{DDIM}} > \tau$, indicating insufficient purification, $\boldsymbol{z}_0^{k+1}$ is sent to the next epoch for further inversion optimization. Otherwise, if $\mathcal{L}_{\text{DDIM}} \leq \tau$, the latent is considered purified, and $\boldsymbol{z}_0^{k+1}$ is decoded by the pretrained VAE decoder to produce the final purified image $\boldsymbol{x}'$.

Inversion Optimization

As analyzed above, the DDIM inversion reconstruction error serves as a reliable metric for distinguishing between clean and adversarial images. Building on this insight, we propose the DDIM metric loss $\mathcal{L}_{\text{DDIM}}$ as an indicator of adversarial images by the DDIM reconstruction. By minimizing the DDIM metric loss $\mathcal{L}_{\text{DDIM}}$, inversion optimization process can remove adversarial noise from the latent representation $\boldsymbol{z}_0^k$. Specifically, leveraging a pretrained UNet $\epsilon_{\boldsymbol{\theta}}$, the latent $\boldsymbol{z}_0^k$ is first sampled to timestamp $\hat{T}$ through the DDIM inversion process $q_{\boldsymbol{\theta}}$ as:

$$q_{\boldsymbol{\theta}}\left(\boldsymbol{z}_{1:\hat{T}}^k | \boldsymbol{z}_0^k\right) = \prod_{t=1}^{\hat{T}} q_{\boldsymbol{\theta}}(\boldsymbol{z}_t^k | \boldsymbol{z}_{t-1}^k). \tag{8}$$

The inverted latent $z_{\hat{T}}^k$ is then sampled back to timestamp 0 by the DDIM process $p_{\boldsymbol{\theta}}$ as:

$$p_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{\hat{T}-1:0}^k | \hat{\boldsymbol{z}}_{\hat{T}}^k) = \prod_{t=1}^{\hat{T}} p_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{t-1}^k | \hat{\boldsymbol{z}}_t^k), \tag{9}$$

where $\hat{\boldsymbol{z}}_{\hat{T}}^k = \boldsymbol{z}_{\hat{T}}^k$ as an initialization. The reconstructed image $\hat{\boldsymbol{x}}$ is obtained by decoding $\hat{\boldsymbol{z}}_0^k$ with the pretrained VAE decoder as $\hat{\boldsymbol{x}} = \mathcal{D}(\hat{\boldsymbol{z}}_0^k)$. The DDIM metric loss, which quantifies the distance between the reconstructed image $\hat{\boldsymbol{x}}$ and the input image $\boldsymbol{x}$, is computed using the loss function $\mathcal{L}_{\text{DDIM}}$:

$$\mathcal{L}_{\text{DDIM}} = \left\| \mathcal{D}\left(p_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{z}}_{\hat{T}-1:0}^k | q_{\boldsymbol{\theta}}\left(\boldsymbol{z}_{1:\hat{T}}^k | \boldsymbol{z}_0^k \right) \right) \right) - \boldsymbol{x} \right\|_2^2. \tag{10}$$

Thus, the inversion optimization problem $P.1$ of z_0^k can be formally defined as:

$$P.1 \quad \min_{\boldsymbol{z}_0^k} \mathbb{E}_{\epsilon_{\boldsymbol{\theta}}, \boldsymbol{z}_0^0 = \mathcal{E}(\boldsymbol{x})} \left[\mathcal{L}_{\text{DDIM}}\left(\boldsymbol{x}, \boldsymbol{z}_0^k, \boldsymbol{c}, \hat{T} \right) \right]. \tag{11}$$

In summary, this optimization process iteratively refines the latent representation $\boldsymbol{z}_0^k$ to minimize the DDIM metric loss, effectively removing adversarial noise while preserving the content of the image.

Dynamic Optimization Epochs

In practical scenarios, datasets often contain a mixture of adversarial and clean images. Applying the same number of optimization epochs to all images would be computationally inefficient, especially when many images are already clean or require minimal purification. To address this, we introduce a dynamic optimization epochs strategy, which adaptively adjusts the number of optimization epochs based on the DDIM metric loss of each image. The key idea is to set a tradeoff threshold τ that determines when the optimization process should terminate. If the DDIM metric loss $\mathcal{L}_{\text{DDIM}}$

Attacks	Purification	FDFR↓	ISM↑	SER-FQA↑	BRISQUE↓	FID↓	NIQE↓
PID	UDAP	**0.14**	**0.57**	**0.59**	**21.45**	**185.72**	**4.38**
	DiffPure	0.24	0.42	0.42	29.93	238.52	4.54
	GridPure	0.21	0.48	0.46	25.29	205.62	4.46
	-	0.87	0.02	0.06	48.24	414.35	5.96
MIST	UDAP	0.11	**0.61**	**0.73**	**20.42**	**163.25**	**4.24**
	DiffPure	**0.10**	0.56	0.67	25.64	224.33	4.53
	GridPure	0.11	0.58	0.66	22.83	185.72	4.35
	-	0.12	0.51	0.63	33.72	273.37	4.93
Anti-DB	UDAP	**0.09**	**0.62**	**0.72**	**17.53**	**142.54**	**4.21**
	DiffPure	0.14	0.58	0.66	26.48	205.82	4.48
	GridPure	0.11	0.58	0.69	22.75	158.54	4.74
	-	0.55	0.40	0.38	38.24	342.36	5.58
Anti-DF	UDAP	**0.22**	**0.51**	**0.64**	**25.08**	**202.67**	**4.56**
	DiffPure	0.46	0.33	0.55	29.50	356.22	4.57
	GridPure	0.33	0.46	0.56	27.32	235.11	4.62
	-	0.59	0.25	0.45	39.46	257.98	5.82
MetaCloak	UDAP	**0.24**	**0.53**	**0.58**	**30.36**	**264.81**	**4.49**
	DiffPure	0.54	0.31	0.44	35.15	332.52	4.62
	GridPure	0.48	0.35	0.51	34.52	325.74	4.64
	-	0.73	0.03	0.08	47.38	404.82	5.88
clean	-	0.09	0.63	0.74	18.36	142.38	4.34

Table 1: Comparison on the purification performance of different methods on the DreamBooth model on dataset CelebA-HQ. The best-performing purification under each metric is marked with **bold**.

for a given latent $\boldsymbol{z}_0^{k+1}$ exceeds τ, indicating that the image is not yet sufficiently purified, the latent will be sent to the next epoch for further optimization. The optimization continues until either $\mathcal{L}_{\text{DDIM}} \leq \tau$ or the maximum epoch K is reached. This threshold should be able to represent the average performance of clean images in DDIM inversion reconstruction. Therefore, we set the threshold τ as the average DDIM metric loss on a batch of N clean images, which can be estimated as:

$$\tau \approx \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{\text{DDIM}}(\boldsymbol{x}_n, \epsilon_{\boldsymbol{\theta}}, c, \hat{T}). \tag{12}$$

In our UDAP, τ is estimated by 1000 clean images from ImageNet (Deng et al. 2009) as $\tau = 4 \times 10^{-3}$. The impact of different τ values on the purification performances and efficiency of UDAP will also be discussed in subsequent ablation studies. By dynamically adjusting the number of optimization epochs based on the DDIM metric loss, UDAP achieves a balance between purification quality and computational efficiency. This strategy ensures that heavily adversarial perturbed images undergo sufficient optimization, while clean or minimally perturbed images are processed efficiently, making the framework highly practical for real-world applications.

Experiment

Implementation Details

Datasets. To conduct adversarial attacks and purifications and to train the DreamBooth models, we utilize the dataset methodology from Anti-DB. Specifically, we conduct experiments using 100 unique identifiers (IDs) sourced from the VGGFace2 (Cao et al. 2018) and CelebA-HQ (Karras et al. 2017) datasets where each ID contains 4 images.

Training Configurations. During the purification process of UDAP, the threshold τ is 4×10^{-3} and the max epoch number K is 100 by default. To balance the consumption of GPU memory and the precision of DDIM Inversion, we set the total inference steps T and max depth $\hat{T}$ for DDIM inversion to 20 and 10, respectively. Null prompt is used in the DDIM inversion. The entire purification process, when executed on 8 NVIDIA A6000 GPUs, takes approximately 20 minutes for 400 images with the shape of 512×512 (around 3 seconds per image).

Evaluation Metrics. We perform purification on adversarial images using various adversarial attack methods. Subsequently, we fine-tune SD with DreamBooth to evaluate the purification performance. To measure the purification performance on the DreamBooth models, following Anti-DB, we adopt the following four metrics: BRISQUE (Mittal, Moorthy, and Bovik 2012), SER-FQA (Terhorst et al. 2020), FDFR (Deng et al. 2020), and ISM (Deng et al. 2019). Additionally, we introduce two more Image Quality Assessment (IQA) metrics: Fréchet Inception Distance (FID) (Heusel et al. 2017) and Natural Image Quality Evaluator (NIQE) (Mittal, Soundararajan, and Bovik 2012).

Comparison with Purification Baselines

We evaluate UDAP's performance against a diverse set of adversarial attacks, including VAE-targeted (PID), UNet-targeted (Anti-DB), and mixed-strategy (MIST) methods. We also tested our UDAP on robust adversarial attack methods (Anti-DF and MetaCloak) that are specially designed to attack purification methods. For baselines, though most

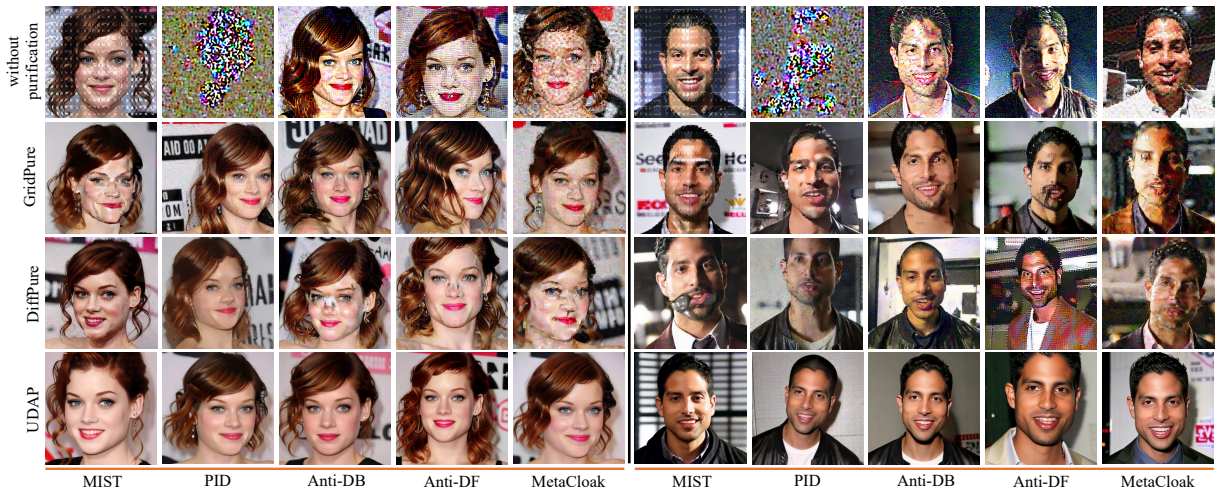

Figure 4: Qualitative purification results of different methods against different adversarial attacks on the DreamBooth model. The specific prompt adopted is "a photo of sks person". The instances are from CelebA-HQ (left) and VGGFace2 (right).

adversarial purification methods are targeting classification, we also test recent diffusion based general purification methods, like DiffPure (Nie et al. 2022) and GridPure(Zhao et al. 2024), as a comparison. All these methods are processed in their default configurations. During the evaluation process, for each trained DreamBooth model, we generate 16 images under 5 different seeds, totaling 80 images, to evaluate the corresponding results, thereby eliminating the variability associated with a single seed.

The quantitative results on DreamBooth for purification of different adversarial methods are shown in Tab. 1. Compared to clean images, adversarial attacks severely disrupt DreamBooth's performance; for instance, the MetaCloak causes the identity-preserving ISM score to plummet from 0.63 to 0.03. After applying various purification methods, performance improves, but UDAP consistently achieves the best results across most metrics. Against the Anti-DB, UDAP-purified images produce generations with superior facial integrity, achieving the lowest FDFR (0.09) and highest SER-FQA (0.72). It also best recovers the image's ID features, reaching the highest ISM value of 0.57 against PID, a significant improvement over baselines like DiffPure (0.42). Additionally, UDAP delivers the best image quality, achieving the lowest FID score of 264.81 against the robust MetaCloak attack, whereas other methods score above 325. In summary, for images from CelebA-HQ, UDAP provides superior adversarial purification performance. More results on VGG-Face can be found in the supplementary.

The qualitative results in Fig. 4 further support that UDAP provides superior adversarial purification performance. While methods like GridPure and DiffPure offer some level of purification by promoting the visual quality of generated images, UDAP could purify most adversarial perturbations and make DreamBooth to generate images with the best visual quality. The advantages of UDAP are very evident, especially in purifying images affected by robust methods such as Anti-DF and MetaCloak, where it can minimize visual loss to the greatest extent. As shown in Fig. 5, UDAP can remove adversarial noise from adversarial images for both VAE targeted method (PID) and UNet targeted method (Anti-DB), while other purification method still leave a lot of visible noise on the purified images.

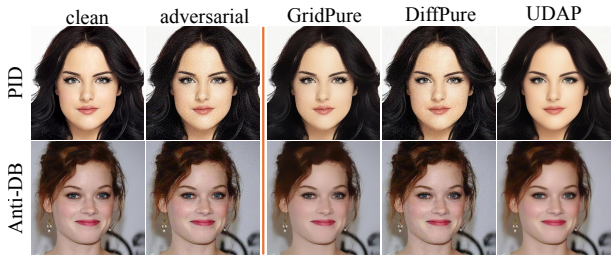

Figure 5: Comparison on the adversarial images and purified images under different adversarial attack and purification methods.

Ablation Study

To dynamically adjust the number of inversion optimization epochs, a preset threshold τ is used to determine whether the latent z_0 images are well purified. To analyze the impact of τ on the purification performance and efficiency of UDAP, we conduct comparative experiments with different values of τ on DreamBooth against Anti-DB. We constructed a dataset based on CelebA-HQ consisting of 50 unique IDs, with 4 images per ID. To simulate real-world scenarios where clean and adversarial images are mixed, 2 images per ID are adversarial perturbed using Anti-DB, while the remaining 2 images are kept clean. The experimental results are presented in Tab. 2.The first four rows show the purification performance for τ ranging from 2×10^{-3} to 5×10^{-3}. As τ increases, both the purification performance and the average processing time decrease. Notably, $\tau = 4 \times 10^{-3}$ strikes an optimal balance, offering strong purification performance

while maintaining a relatively short processing time.

For comparison, the last two rows in Tab. 2 show the purification performance when using fixed numbers of epochs (100 and 50). While reducing the number of epochs decreases the average processing time, it also leads to a significant drop in purification performance. These results highlight the effectiveness of UDAP's dynamic optimization strategy, which adaptively adjusts the number of epochs based on the DDIM metric loss, ensuring both efficiency and high-quality purification.

value of τ	ISM↑	BRISQUE↓	FID↓	time(s)
2×10^{-3}	0.66	18.33	146.72	18
3×10^{-3}	0.66	18.68	144.81	7
4×10^{-3}	0.63	19.82	145.66	3
5×10^{-3}	0.28	24.54	232.91	2
epochs=100	0.68	18.13	144.27	15
epochs=50	0.22	21.10	164.53	8

Table 2: Comparison of purification performance on DreamBooth against Anti-DB with different values of τ. The last two rows represent results with fixed epochs (100 and 50). The "time" column indicates the average time cost per image. The dataset is derived from CelebA-HQ, where half of the images are adversarial, and the remaining half are clean.

To further assess the impact of UDAP on clean images, we conduct a comparative evaluation between "adv." and "clean" datasets using DreamBooth against Anti-DB. The datasets are constructed from CelebA-HQ, with each identity represented by four images. Here, "adv." refers to datasets composed entirely of adversarial examples, while "clean" denotes datasets containing only clean images. As shown in Tab. 3, the experimental results demonstrate negligible performance differences between clean images processed with and without UDAP. This indicates that UDAP has minimal effect on the behavior of clean images.

UDAP	dataset	ISM↑	BRISQUE↓	FID↓
✓	adv.	0.62	17.53	142.54
✓	clean	0.62	18.41	144.47
×	adv.	0.40	38.24	342.36
×	clean	0.63	18.36	142.38

Table 3: Comparison on the purification performance of UDAP on DreamBooth against Anti-DB under datasets with different degrees of adversarial perturbations.

Unexpected Scenarios

In practical scenarios, the specific utilization of the SD models by malicious users to perform adversarial attacks is unpredictable. We perform experiments to evaluate the purification performance of UDAP across different SD versions when they are match or mismatch. Specifically, we attacks the CelebA-HQ dataset using Anti-DB based on SD v1.4 and v2.1, respectively. We then purified the adversarial images using UDAP based on different SD versions. Subsequently, we trained SD using purified images with the DreamBooth method and analyzed the quality of the generated images. The experimental results are shown in Tab. 4, which demonstrates that UDAP can effectively purify Anti-DB across SD versions, even when the SD versions used for adversarial attacks and purification do not match.

Adv.	Pur.	FDFR↓	ISM↑	BRISQUE↓	FID↓
v2.1	v2.1	0.09	0.62	17.53	142.54
	v1.4	0.10	0.61	18.73	144.78
	no	0.55	0.40	38.24	342.36
v1.4	v2.1	0.09	0.63	17.24	146.79
	v1.4	0.11	0.64	16.37	141.29
	no	0.54	0.38	38.36	332.76

Table 4: Comparison on the purification performance of UDAP against Anti-DB under different versions of SD on dataset CelebA-HQ. The terms "Adv." and "Pur." refer to the SD version for adversarial attacks with Anti-DB and purifying with UDAP.

For DreamBooth, different prompts can be used to generate various content. To investigate the impact of different prompts on the purification effect of UDAP against Anti-DB, we additionally introduce three prompts: p1 , p2, and p3, which are "a photo of sks person with sad face", "facial close-up of sks person" and "a photo of sks person yawning in speech" respectively, to evaluate the purification performance. We can see from Tab. 5 that UDAP can also provide purification across different prompts.

P	Pur.	FDFR↓	ISM↑	BRISQUE↓	FID↓
p1	yes	**0.11**	**0.53**	**17.28**	**183.39**
	no	0.44	0.32	37.52	346.2
p2	yes	**0.07**	**0.44**	**16.83**	**153.82**
	no	0.62	0.13	29.51	322.71
p3	yes	**0.09**	**0.33**	**19.72**	**201.44**
	no	0.67	0.11	36.77	412.52

Table 5: Comparison on the purification performance with different inference prompts on dataset CelebA-HQ. "P" and "Pur." refer to prompt and purification.

Conclusion

This paper presents UDAP, a universal adversarial purification framework specifically designed to mitigate adversarial attacks on SD. UDAP leverages the distinct reconstruction behaviors of clean and adversarial images during DDIM inversion, introducing a DDIM metric loss that effectively eliminates adversarial perturbations while preserving the core content of the input image. UDAP also incorporates a dynamic epoch adjustment strategy, which adaptively optimizes the purification process and significantly improves its computational efficiency. Extensive experimental results demonstrate that UDAP surpasses existing methods in defending against diverse adversarial techniques. Furthermore, UDAP achieves superior purification under cross-version SD scenarios and varying inference prompts, showcasing its generalizability in real-world applications.

Acknowledgments

This work was supported in part by Macau Science and Technology Development Fund under 001/2024/SKL, 0119/2024/RIB2, and 0022/2022/A1; in part by Research Committee at University of Macau under MYRG-GRG2023-00058-FST-UMDF; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012536.

References

Bai, T.; Luo, J.; Zhao, J.; Wen, B.; and Wang, Q. 2021. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*.

Baniecki, H.; and Biecek, P. 2024. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 102303.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; and Mukhopadhyay, D. 2021. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1): 25–45.

Chen, J.; Wu, Y.; Luo, S.; Xie, E.; Paul, S.; Luo, P.; Zhao, H.; and Li, Z. 2024. PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models. arXiv:2401.05252.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426.

Costa, J. C.; Roxo, T.; Proença, H.; and Inácio, P. R. 2024. How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access*.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, 89–106. Springer.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Li, A.; Mo, Y.; Li, M.; and Wang, Y. 2024. PID: Prompt-Independent Data Protection Against Latent Diffusion Models. *arXiv preprint arXiv:2406.15305*.

Li, X.; Xin, Z.; and Liu, W. 2022. Defending against adversarial attacks via neural dynamic system. *Advances in Neural Information Processing Systems*, 35: 6372–6383.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 20763–20786. PMLR.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1778–1787.

Lin, G.; Tao, Z.; Zhang, J.; Tanaka, T.; and Zhao, Q. 2024. Robust Diffusion Models for Adversarial Purification. *arXiv preprint arXiv:2403.16067*.

Liu, Y.; Fan, C.; Dai, Y.; Chen, X.; Zhou, P.; and Sun, L. 2024. MetaCloak: Preventing Unauthorized Subject-driven Text-to-image Diffusion-based Synthesis via Meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24219–24228.

Meng, D.; and Chen, H. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 135–147.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Samangouei, P. 2018. Defense-gan: protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *Advances in neural information processing systems*, 32.

Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J.; Davis, L. S.; and Goldstein, T. 2020. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5636–5643.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Terhorst, P.; Kolf, J. N.; Damer, N.; Kirchbuchner, F.; and Kuijper, A. 2020. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5651–5660.

Truong, V. T.; Dang, L. B.; and Le, L. B. 2024. Attacks and Defenses for Generative Diffusion Models: A Comprehensive Survey. *arXiv preprint arXiv:2408.03400*.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2116–2127.

Wang, J.; Lyu, Z.; Lin, D.; Dai, B.; and Fu, H. 2022. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*.

Wang, K.; Fu, X.; Han, Y.; and Xiang, Y. 2024. DiffHammer: Rethinking the robustness of diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 37: 89535–89562.

Yang, Z.; Feng, R.; Zhang, H.; Shen, Y.; Zhu, K.; Huang, L.; Zhang, Y.; Liu, Y.; Zhao, D.; Zhou, J.; and Cheng, F. 2024. Lipschitz Singularities in Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Yoon, J.; Hwang, S. J.; and Lee, J. 2021. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, 12062–12072. PMLR.

Zhao, Z.; Duan, J.; Xu, K.; Wang, C.; Zhang, R.; Du, Z.; Guo, Q.; and Hu, X. 2024. Can Protective Perturbation Safeguard Personal Data from Being Exploited by Stable Diffusion? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24398–24407.

Zheng, L.; Xie, L.; Zhou, J.; Wang, X.; Wu, H.; and Tian, J. 2025. Anti-Diffusion: Preventing Abuse of Modifications of Diffusion-Based Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10): 10582–10590.

Zollicoffer, G.; Vu, M. N.; Nebgen, B.; Castorena, J.; Alexandrov, B.; and Bhattarai, M. 2025. Lorid: Low-rank iterative diffusion for adversarial purification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23081–23089.